# Chapter 3

## Section 3.1

### *Check Your Understanding, page 144:*

1. The explanatory variable is the number of cans of beer. The response variable is the blood alcohol level.
2. There are two explanatory variables: amount of debt and income. The response variable is stress caused by college debt.

### *Check Your Understanding, page 149:*

1. The relationship is positive. The longer the duration of the eruption, the longer we should expect to wait between eruptions. One reason for this may be that if the geyser erupted for longer, it expended more energy and it will take longer to build up the energy needed to erupt again.
2. The form is roughly linear with two clusters. The clusters indicate that in general there are two types of eruptions—shorter eruptions that last around 2 minutes and longer eruptions that last around 4.5 minutes.
3. The relationship is fairly strong. The points don't deviate much from the linear form.
4. There are a few possible outliers around the clusters. However, there aren't many and potential outliers are not very distant from the main clusters of points.
5. The Starnes family needs to know how long the last eruption was in order to predict how long it will be until the next one.

### *Check Your Understanding, page 153:*

(a) The correlation is about $r = 0.9$. This indicates that there is a strong, positive linear relationship between the number of boats registered in Florida and the number of manatees killed.
(b) The correlation is about $r = 0.5$. This indicates that there is a moderate, positive linear relationship between the number of named storms predicted and the actual number of named storms.
(c) The correlation is about $r = 0.3$. This indicates that there is a weak, positive linear relationship between the healing rate of the two front limbs of the newts.
(d) The correlation is about $r = -0.1$. This indicates that there is a weak, negative linear relationship between last year's percent return and this year's percent return in the stock market.

***Exercises, page 159:***

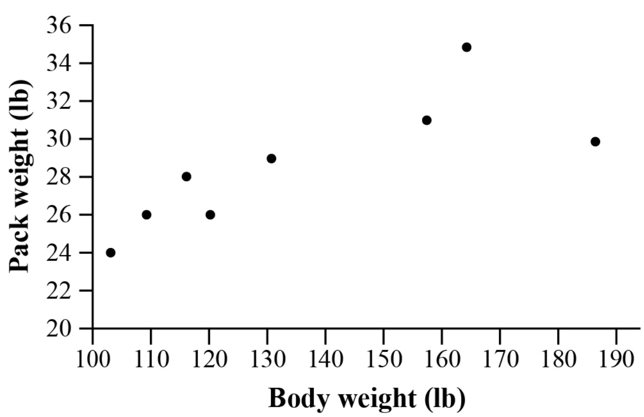
3.1 Water temperature is the explanatory variable and weight change (growth) is the response variable. Both are quantitative.

3.2 The explanatory variable is the type of treatment—removal of the breast or removal of only the tumor and nearby lymph nodes, followed by radiation. Survival time is the response variable. Type of treatment is a categorical variable and survival time is a quantitative variable.

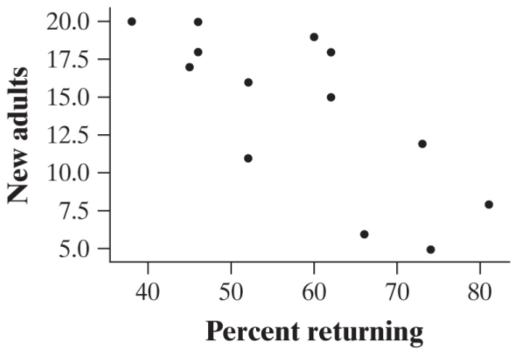
3.3 (a) Positive association. Students with higher IQs tend to have higher GPAs, and those with lower IQs generally have lower GPAs. This makes sense because both IQ and GPA are related to mental ability.
(b) The form of the relationship is roughly linear, because a line through the scatterplot of points would provide a good summary. The association is moderately strong (with a few exceptions) because most of the points would be close to the line.
(c) The lowest point on the plot is for a student with an IQ of about 103 and a GPA of about 0.4.

3.4 (a) Negative association. Months with high temperatures tend to have low gas consumption and months with low temperatures tend to have high gas consumption. This makes sense because as the temperature gets warmer, Joan would not need to use as much gas to heat her house.
(b) The scatterplot shows a strong linear relationship. It is linear because a line through the scatterplot of points would provide a good summary and it is strong because the points would all be close to the line. (c) The point in the bottom right of the plot represents a month when the average temperature was about 58 degrees and the gas usage was about 260 cubic feet.

3.5 A scatterplot is shown below.



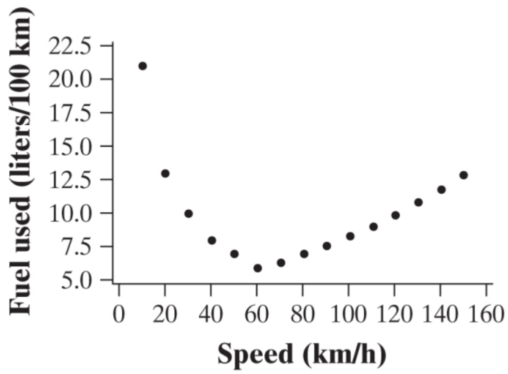
3.6 A scatterplot is shown below.




3.7 (a) *Direction*: There is a positive association between backpack weight and body weight. In general, lighter students are carrying lighter backpacks and heavier students are carrying heavier packs. *Form*: For students under 140 pounds, there seems to be a linear pattern in the graph. However, for students above 140 pounds, the association begins to curve. *Strength*: Because the points vary somewhat from the linear pattern, the relationship is only moderately strong.
(b) We see one possible outlier in the graph—the hiker with body weight 187 pounds and pack weight 30 pounds. This hiker makes the form appear to be nonlinear for weights above 140 pounds. Without this hiker, the association would look very linear for all body weights.

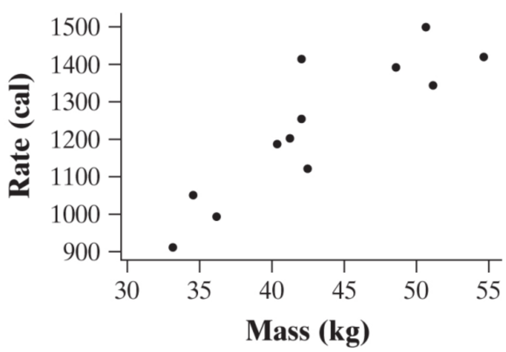
3.8 (a) *Direction*: There is a negative association between percent returning and the number of new adults. In general, when there is a higher percent returning there are fewer new adults. *Form*: The association is linear because a line through the scatterplot of points would provide a good summary. *Strength*: Because the points vary somewhat from the linear pattern, the relationship is only moderately strong.
(b) Because this association is negative, we conclude that the sparrowhawk is a long-lived territorial species.

3.9 (a) A scatterplot with speed as the explanatory variable is shown below.




(b) The relationship is curved. Large amounts of fuel were used for low and high values of speed and smaller amounts of fuel were used for moderate speeds. This makes sense because the best fuel efficiency is obtained by driving at moderate speeds. (Note: 60 km/hr is about 37 mph)
(c) Both directions are present in the scatterplot. The association is negative for lower speeds and positive for higher speeds.
(d) The relationship is very strong, with little deviation from a curve that can be drawn through the points.

3.10 (a) A scatterplot with mass as the explanatory variable is shown below.

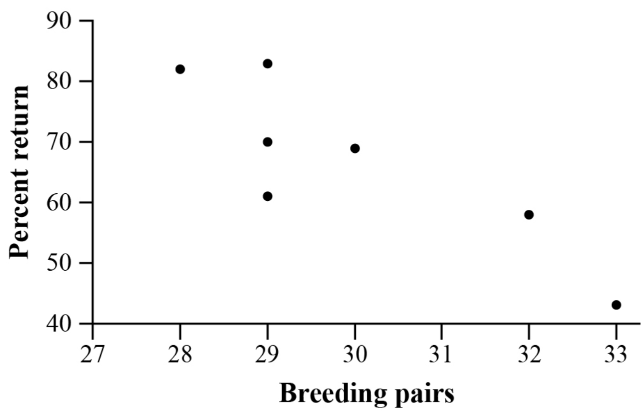



(b) *Direction*: There is a positive association between lean body mass and metabolic rate. In general, people who have more lean body mass tend to have greater metabolic rates. *Form*: The association is linear because a line drawn through the points on the scatterplot would be a good summary of the relationship. *Strength*: The association is strong because the points do not deviate much from the linear form. This association is consistent with the researchers' belief that lean body mass is an important influence on metabolic rate.

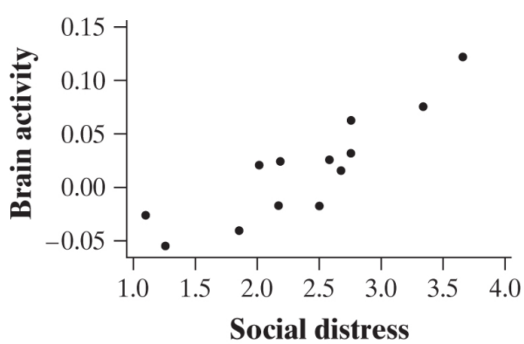
3.11 (a) Most of the southern states fall in the same pattern as the rest of the states. However, the southern states typically appear at the bottom of their clusters. This means that southern states typically have lower mean SAT math scores than other states with a similar percent of students taking the SAT.
(b) West Virginia is an outlier because it has a much lower mean SAT Math score than the other states that have a similar percent of students taking the exam.

3.12 For both men and women there is a relatively strong, positive linear association between lean body mass and metabolic rate. However, the relationship between mass and rate is not as strong for men as it is for women. Also, the men tend to have higher lean body masses and higher metabolic rates than the women.

3.13 A scatterplot displaying the relationship is shown below. The scatterplot of the percent returning against the number of breeding pairs shows the expected negative association—in years where there are more breeding pairs, a smaller percent return. The association is linear and moderately strong, as the points aren't too spread out from the linear form.




3.14 A scatterplot displaying the relationship is shown below. The scatterplot of brain activity against the social distress score shows the expected positive association. Subjects with higher social distress scores tend to have higher amounts of brain activity. The association is linear and strong as the points don't deviate too much from the linear form.



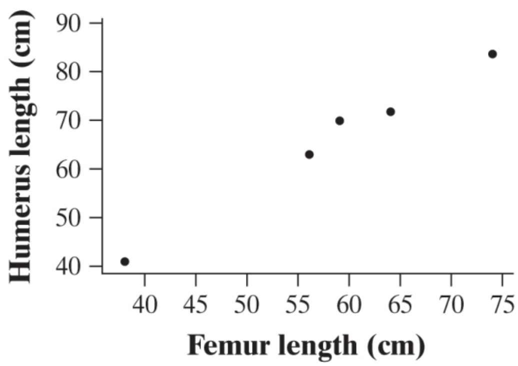

3.15 (a) $r = 0.9$ (b) $r = 0$ (c) $r = 0.7$ (d) $r = -0.3$ (e) $r = -0.9$

3.16 We would expect the height of women at age 4 and their height as women at age 18 to be the highest correlation since it is reasonable to expect taller children to become taller adults and shorter children to become shorter adults. The next highest would be the correlation between the heights of male parents and their adult children because they share genes. Tall fathers tend to have relatively tall sons and short fathers tend to have relatively short sons. The lowest correlation would be between husbands and their wives. Some tall men may prefer to marry tall women, but this isn't always the case.

3.17 (a) Gender is a categorical variable and the correlation $r$ measures the strength of linear association for two quantitative variables.
(b) The largest possible value of the correlation is $r = 1$.
(c) The correlation $r$ has no units.

3.18 The paper's report is wrong because the correlation of $r \approx 0$ means that there is no linear association between research productivity and teaching rating. In other words, knowledge of a professor's research productivity will not help you predict her teaching rating. The author incorrectly suggests that a correlation of zero indicates a negative association between research productivity and teaching rating.

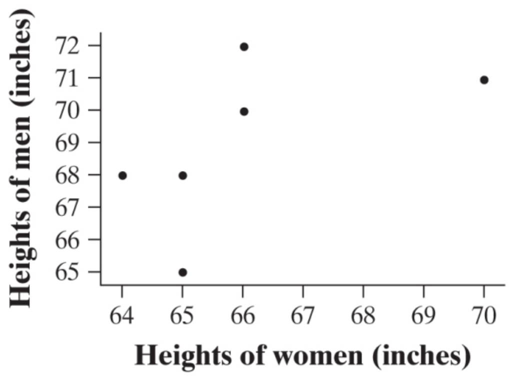
3.19 (a) The scatterplot below shows a strong, positive, linear relationship between the two measurements. It appears that all five specimens come from the same species.




(b) The femur measurements have mean of 58.2 cm and a standard deviation of 13.2 cm. The humerus measurements have a mean of 66 cm and a standard deviation of 15.89 cm. The table below shows the standardized measurements (labeled zfemur and zhumerus) obtained by subtracting the mean and dividing by the standard deviation. The column labeled "product" contains the product (zfemur×zhumerus) of the standardized measurements. The sum of the products is 3.97620, so the correlation coefficient is $r = \frac{1}{4}(3.97620) = 0.9941$. The very high value of the correlation confirms the strong, positive linear association between femur length and humerus length in the scatterplot from part (a).

| Femur | Humerus | zfemur | zhumerus | product |
|---|---|---|---|---|
| 38 | 41 | –1.53030 | –1.57332 | 2.40765 |
| 56 | 63 | –1.16667 | –0.18880 | 0.03147 |
| 59 | 70 | 0.06061 | 0.25173 | 0.01526 |
| 64 | 72 | 0.43939 | 0.37760 | 0.16591 |
| 74 | 84 | 1.19697 | 1.13279 | 1.35591 |

3.20 (a) The scatterplot shows a moderate positive association, so $r$ should be positive, but not close to 1.




(b) For the women, the mean is 66 inches and the standard deviation is 2.098 inches. For the men, the mean is 69 inches and the standard deviation is 2.53 inches. The table below shows the standardized measurements (labeled zfemale and zmale) obtained by subtracting the mean and dividing by the standard deviation. The column labeled "product" contains the product (zfemale×zmale) of the standardized measurements. The sum of the products is 2.82596, so the correlation coefficient is $r = \frac{1}{5}(2.82596) = 0.5652$. Because the correlation is positive, it provides some evidence that taller women tend to date taller men (and shorter women date shorter men). However, because the correlation is not close to 1, the association between heights of females and their boyfriends isn't that strong.

| female | male | zfemale | zmale | product |
|---|---|---|---|---|
| 66 | 72 | 0 | 1.18577 | 0 |
| 64 | 68 | –0.95329 | –0.39526 | 0.37679 |
| 66 | 70 | 0 | 0.39526 | 0 |
| 65 | 68 | –0.47664 | –0.39526 | 0.18840 |
| 70 | 71 | 1.90658 | 0.79051 | 1.50718 |
| 65 | 65 | –0.47664 | –1.58103 | 0.75359 |

3.21 (a) There is a strong, positive linear association between sodium and calories. High-calorie hot dogs tend to be high in sodium, and low-calorie hot dogs tend to be low in sodium.
(b) The hot dog with the lowest calorie content increases the correlation. It falls in the linear pattern of the rest of the data and observations with unusually small or unusually large values of $x$ have a big influence on the correlation.

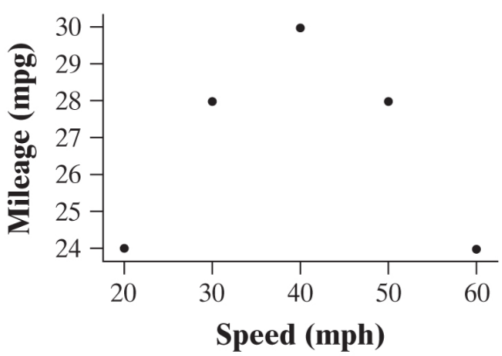
3.22 (a) There is a strong, positive linear association between body weight and brain weight of mammals.

(b) The elephant increases the correlation. It falls in the linear pattern of the rest of the data and observations with unusually small or unusually large values of $x$ have a big influence on the correlation.

3.23 (a) The correlation would not change, because correlation is not affected by a change of units for either variable. In terms of the formula for correlation, multiplying both the $x$ and $y$ values by 10 will also multiply their standard deviations by 10, so the $z$-scores will not change.
(b) The correlation would not change. The correlation measures the strength of the linear relationship between two quantitative variables. It does not distinguish between the explanatory and response variables. In terms of the formula for correlation, reversing the $x$ and $y$ values won't change the correlation because $z_x z_y$ is the same as $z_y z_x$.

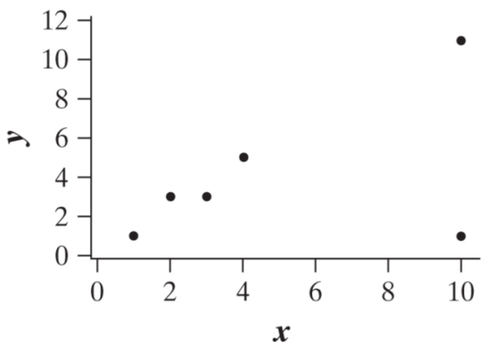
3.24 (a) If all the men were 6 inches shorter, the correlation would not change. In terms of the formula for correlation, subtracting 6 from each $y$-value would also subtract 6 from the mean of $y$, so the $z$-scores would remain unchanged. The correlation tells us that there is a weak to moderate association between women's heights and men's heights (that is, that taller women tend to date taller men), but it does not tell us whether or not women tend to date men taller than themselves.
(b) The correlation would not change, because correlation is not affected by a change of units for either variable. In terms of the formula for correlation, multiplying both the $x$ and $y$ values by 2.54 will also multiply their standard deviations by 2.54, so the $z$-scores will not change.

3.25 (a) A scatterplot of mileage versus speed is shown below.




(b) The correlation is $r = 0$.
(c) The correlation measures the strength of a *linear* association between two quantitative variables. This plot shows a nonlinear relationship between speed and mileage.

3.26 (a) The scatterplot is shown below.




(b) The correlation is $r = 0.481$.
(c) The outlier at (10, 1) is responsible for reducing the correlation. Outliers tend to have a fairly strong effect on correlation, especially when there aren't many other observations.

3.27 a

3.28 e

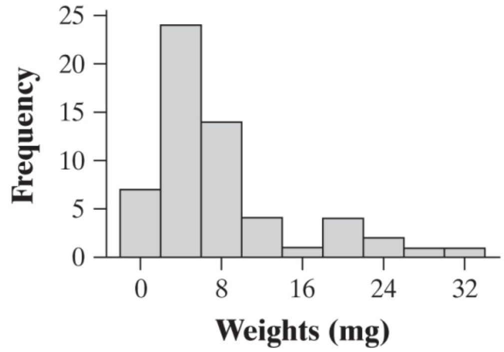
3.29 d

3.30 b

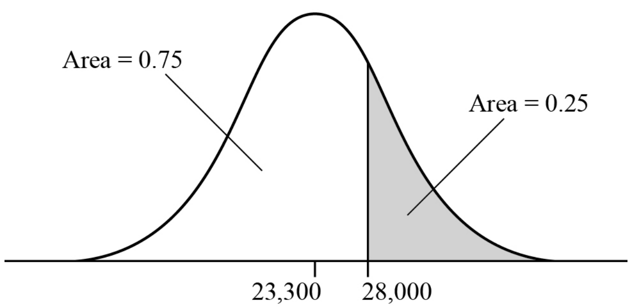
3.31 b

3.32 d

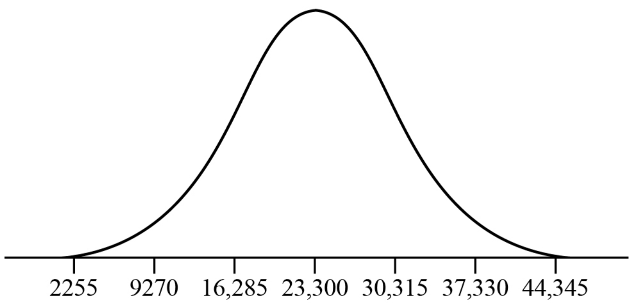
3.33 A histogram is shown below. The distribution is right-skewed, with several possible high outliers. Because of the skewness and outliers, we should use the median (5.4 mg) and *IQR* (5.5 mg) to describe the center and spread.




3.34 (a) Because 25% of borrowers owe more than \$28,000, we need to find the value of $z$ that has 75% of the area to its left (see graph below). Using Table A, this value is $z = 0.67$. To estimate the standard deviation, we need to solve the equation $0.67 = \dfrac{28{,}000 - 23{,}300}{\sigma}$ for $\sigma$. The standard deviation is approximately $\sigma$ = \$7015.




(b) **Step 1: State the distribution and values of interest.** For borrowers, loan balances follow a Normal distribution with mean \$23,300 and standard deviation \$7015. We want to find the proportion of borrowers who owe more than \$54,000 (see graph below). **Step 2: Perform calculations. Show your work.** The standardized score for the boundary value is $z = \dfrac{54{,}000 - 23{,}300}{7015} = 4.38$. From Table A, the proportion of $z$-scores above 4.38 is approximately 0. *Using technology:* The command normalcdf(lower: 54000, upper: 100000, μ: 23300, σ: 7015) gives an area of approximately 0. **Step 3: Answer the question.** Approximately 0% of borrowers owe more than \$54,000.




(c) If the distribution of loan balances is approximately Normal, then we would expect almost no one to have a balance that large. Because 10% of borrowers owe more than $54,000 we can conclude that the distribution of loan balances isn't Normal and is right skewed.
(d) Yes. Because the mean ($23,300) is so much larger than the median ($12,800), we can conclude that the distribution of loan balances is skewed to the right.